Hallucination Rates in 7B versus 70B Language Models on Domain-Specific QA Benchmarks

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does the hallucination rate of 7B parameter models compare to 70B models when evaluated on domain-specific QA benchmarks like TriviaQA or HotpotQA with controlled retrieval context density. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: THaMES: An End-to-End Tool for Hallucination Mitigation and Evaluation in Large Language Models. Research question: How does the hallucination rate of 7B parameter models compare to 70B models when evaluated on domain-specific QA benchmarks like TriviaQA or HotpotQA with controlled retrieval context density?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

16 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.8/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The THaMES framework consists of three main components: testset generation from a user-provided corpus, baseline metric	×	0.05
The THaMES framework transforms raw corpora into specialized LLM benchmarks to reveal optimal hallucination mitigation s	×	0.11
The QA-set generation process in THaMES includes seven specific steps: Knowledge Base Processing, Ground-Truth Weighted	×	0.09
Each QA pair generated by THaMES consists of a question, the correct answer, and one hallucinated answer.	×	0.05
The experimental QA testset for THaMES was generated using a mix of political news articles, academic papers, and Wikipe	×	0.08
The THaMES framework is compatible with PDF, TXT, and CSV file formats.	×	0.02
THaMES utilizes the VectorStoreIndex module provided by LlamaIndex to build a knowledge base from the raw corpus.	×	0.04
The Reasoning Question Evolution Prompt requires modified questions to include at least one leap of intuition to correla	×	0.05
In the Reasoning Question Evolution Prompt, the question_type JSON parameter is modified to 'reasoning' for all evolved	×	0.03
Generated questions must be self-contained, explicitly stating any necessary titles, names, or terms without relying on	×	0.03

References

- http://arxiv.org/abs/2409.11353v3
- http://arxiv.org/abs/2502.12372v1
- http://arxiv.org/abs/2604.00715v1